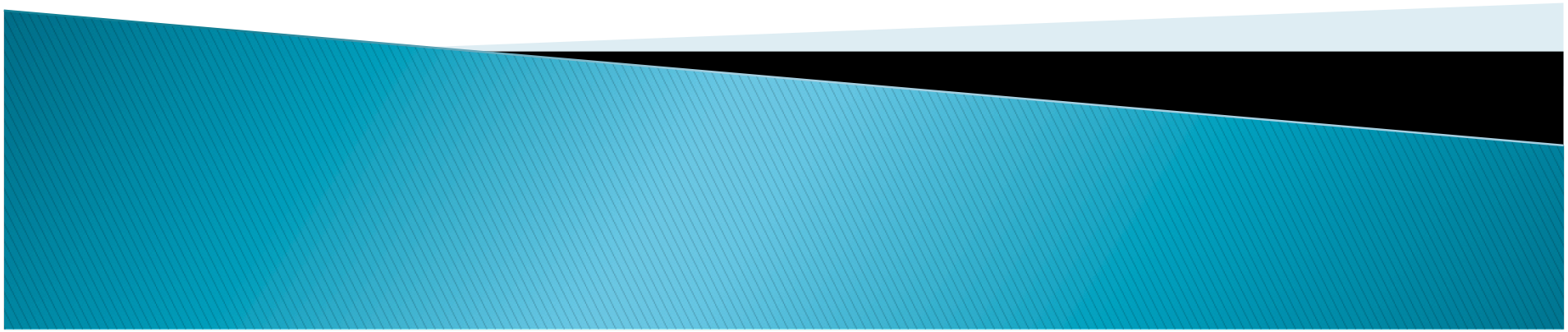
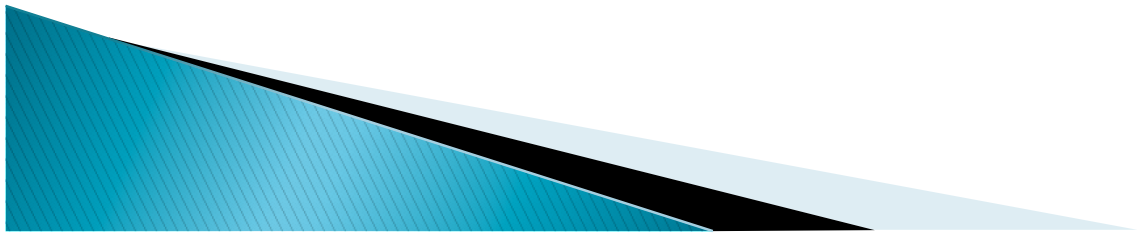
Zeus User Meeting

February 23rd, 2012



Agenda

- Status
- Data Movement
- Outstanding Issues
- New Software Installs
- Tips
- Tuning MPI
- Open Forum



Status

- 215 users have been successfully on-boarded
  - While there are a few outstanding issues, the on-boarding process has smooth over the last few weeks
  - Most issues are due to name changes or initial typos in user records
- Updates to user accounts (project associations are moving along slowly)
  - Over time we will develop a more streamlined process over managing accounts
- Most of the issues are related to the batch system
  - More details to follow

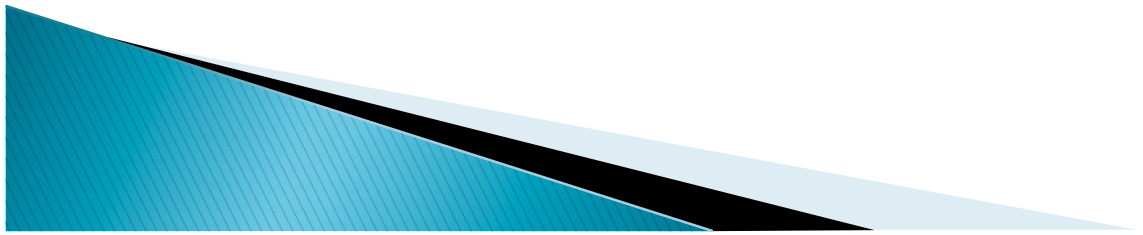
Data Movement

- SSH Port tunnel is working
  - Default method, but not the best method for transferring data
- DTN works, but access is currently limited
  - It does work from DSRC, GFDL (but no DNS yet).
  - NCEP, AOML, NSSL are next in line.
  - We are waiting on the NOC to fix DNS entries before expanding access.
    - We hope to have this resolved in a few days.

NOTE: Please do not use HPSS as a way to transfer data between the CCS and Vapor to Zeus. The DTN does work from there.

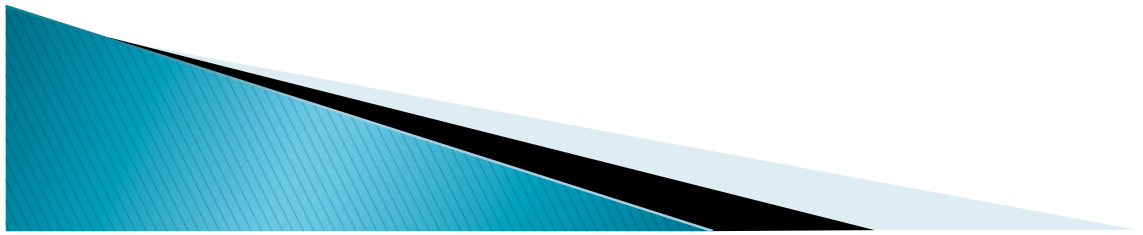
Outstanding Issues

- Details of issues and current status can be found at:
  - https://nesccdocs.rdhpcs.noaa.gov/wiki/index.php/Known_Issues



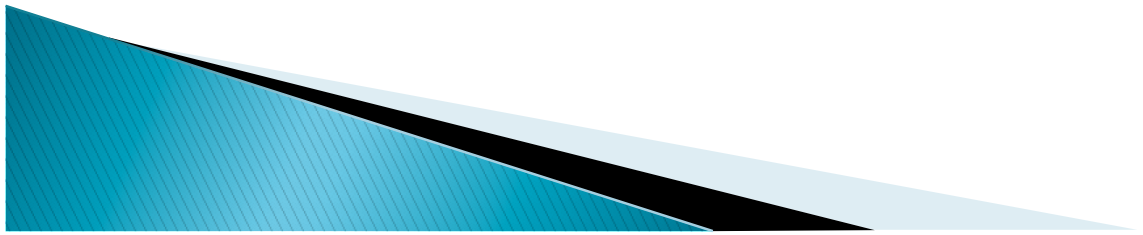
Outstanding Issues, Node Problems

- Processes are getting stuck on the nodes
  - Causes failures of the prologue, epilogue, and job launching
  - Causes jobs to not be cancelled
- SGI has provided some recommendations to resolving this issue



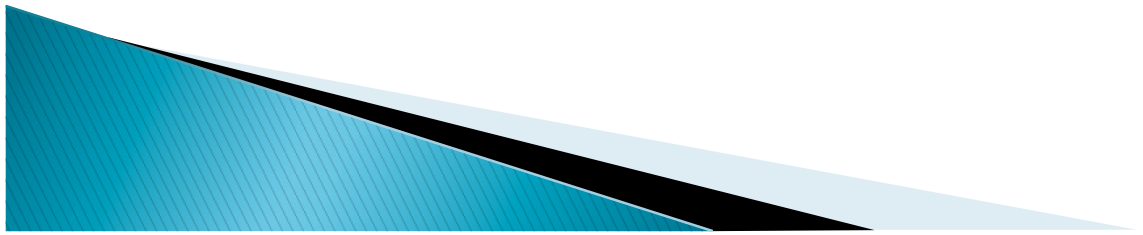
Outstanding Issues, Batch Systems

- Job ID's are Moab.$NUM instead of $NUM
  - Moab scheduler loses contact with Batch server. Vendor has provided some suggestions and will have access to system starting tomorrow to debug
- During msub, "JobCount is 0" is returned, just do run
- PBS_NODEFILE Not found



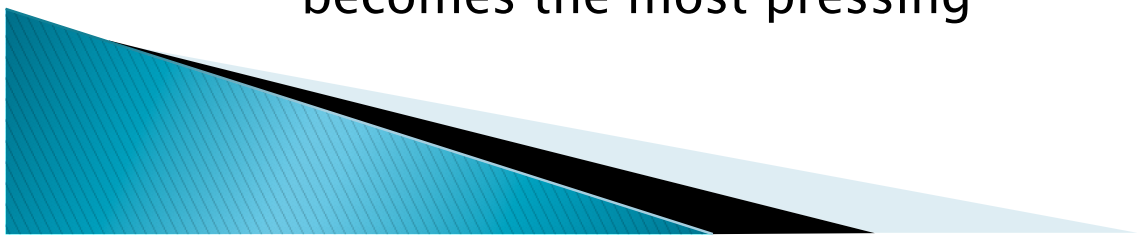
Outstanding Issues, Batch Systems, Cont.

- Mpiexec_mpt cannot find binaries (omplace, executables on $PATH)
  - Bug in Torque. We have a fix to test.
- Msub, mjobctl, canceljob take 1-2 minutes to complete.
  - We have a working theory based on how the server threads. Vendor will start debugging



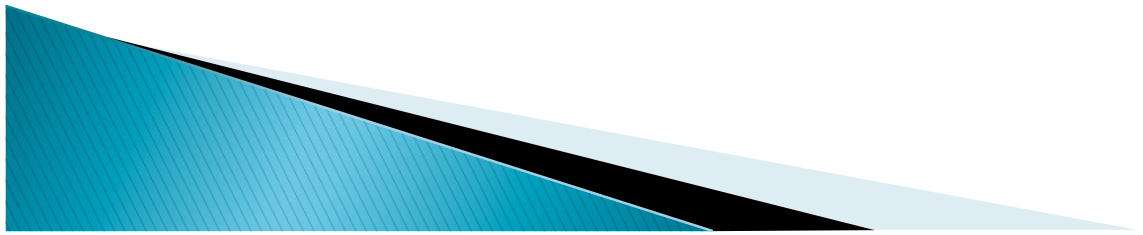
Outstanding Issues, X traffic

- X applications are slow to start
  - Primary issue is speed of light. X traffic is sensitive to latency.
- Several plans of attack for this
  - Use FreeNX, which will bulk transfer data and remove latency sensitivity
  - Experiment and configure quality of service over network to try and prioritize different types of network traffic
    - Ex:
      - ssh or X to front ends, high-priority
      - Other protocols or traffic to DTN, low priority
    - There will be some challenges with this when users use scp from front ends, and we will deal with each bottleneck is it becomes the most pressing



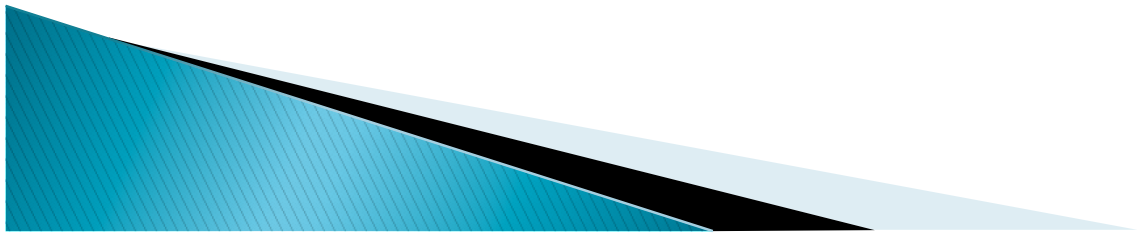
Outstanding issues, other

- Jobs fail with xpmem errors
  - Vendor has been contacted to explain this error and resolve the issue
- “File not found” errors where filename has pre-pended slash
  - No understanding of the problem yet, plan is to reproduce case, try different compiler, and contact vendor.



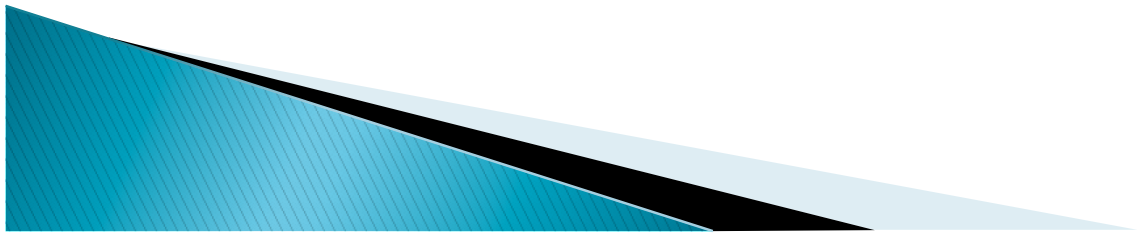
New software installs

- We are working to install new software to meet user needs
- We cannot support all requests, as we need to ensure that the tools will support the larger community not a single user
- Any new requests that have been approved still may take a few weeks as we are trying to stabilize the existing software first



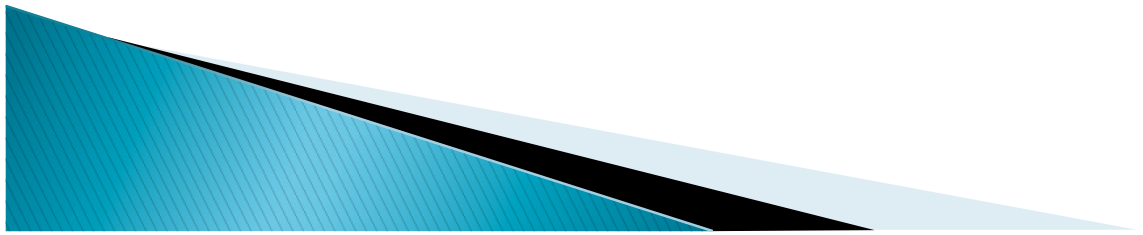
Tips

- Use -A to specify your account when submitting jobs.
  - This feature is not currently working as desired. We will be making changes to require its use.
  - We will give you some warning when we make the change, but not much.
  - Use account_params to find to which projects you are assigned.
- Why is my job not running?
  - Run *checkjob -v* to see



Tuning MPI

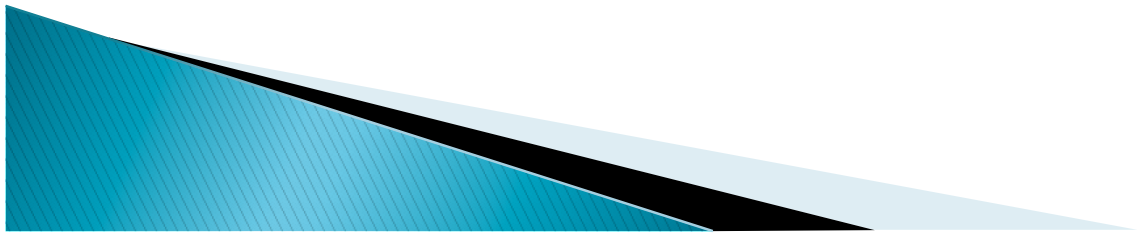
- The default MPI stack, MPT, has been shown in many cases (but not all) to be the fastest
- The problem is that the defaults are low, and your performance will suffer if the buffers are not tuned.
- For details see:
  - https://nesccdocs.rdhpcs.noaa.gov/wiki/index.php/Tuning_MPI



Tuning MPI, Continued

- A better set of buffers to start (ex for bash,ksh):

```
export MPI_BUFS_PER_PROC=128
export MPI_BUFS_PER_HOST=128
export MPI_IB_RAILS=2
export MPI_GROUP_MAX=128
```



MPI Tuning, continued

- How do I know if they are adequate?
  - Profile your application. Set the following in your job:

```
export MPI_STATS=1
export MPI_STATS_FILE=`pwd`/mpi_stats.$PBS_JOBID
```

  - Output will look like:

```
0:0 0 retries allocating mpi PER_PROC headers for collective calls
0:0 0 retries allocating mpi PER_PROC headers for point-to-point calls
0:0 0 retries allocating mpi PER_PROC buffers for collective call
0:0 0 retries allocating mpi PER_HOST buffers for collective calls
0:0 2143 retries allocating mpi PER_PROC buffers for point-to-point call
0:0 160297 retries allocating mpi PER_HOST buffers for point-to-point calls
```

  - Increase the parameters MPI_BUFS_PER_PROC and MPI_BUFS_PER_HOST until there are no retries.

Thank You!

now…
Open Forum